Death of Disks Panel: A Darwinian Evolution

Principles of Operation for Shingled Disk Devices

HEC FSIO 2011, Arlington VA

August 10, 2011

[CMU-PDL-11-107]

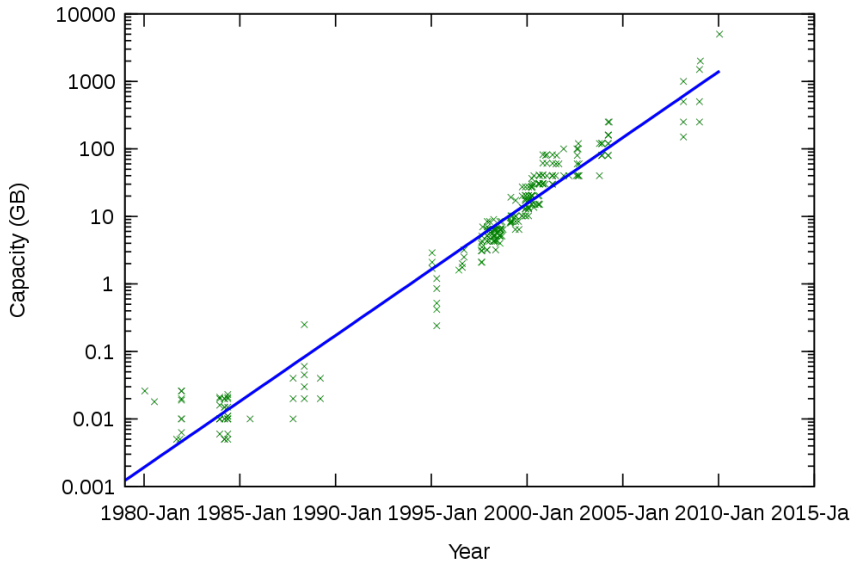
Garth Gibson, Greg Ganger

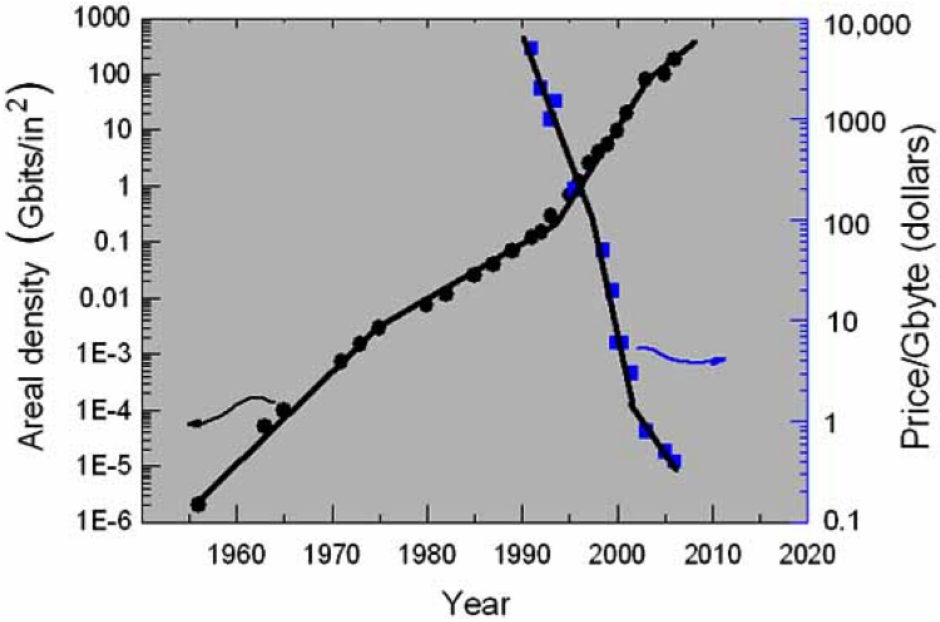
Parallel Data Laboratory, Carnegie Mellon University

garth@cs.cmu.edu, ganger@ece.cmu.edu

Kryder's Law for Magnetic Disks

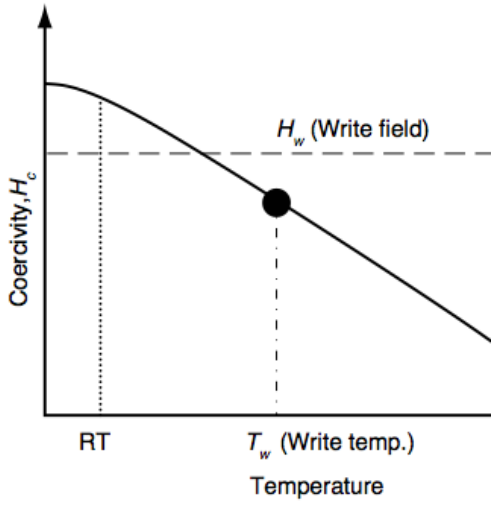
- Market expects ever more dense disks
- Future is multi-terabit per square inch
- Real challenge is making money at $100/disk when engineering is this hard

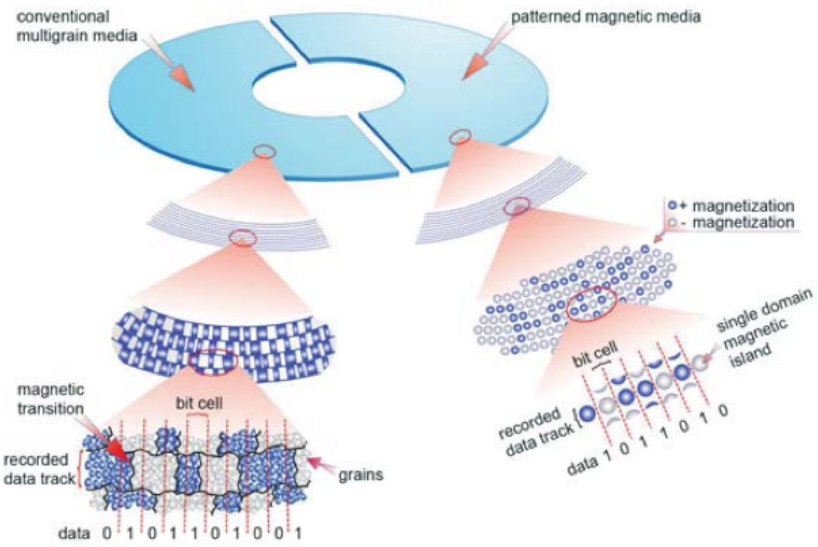




Directions in High Capacity Disks

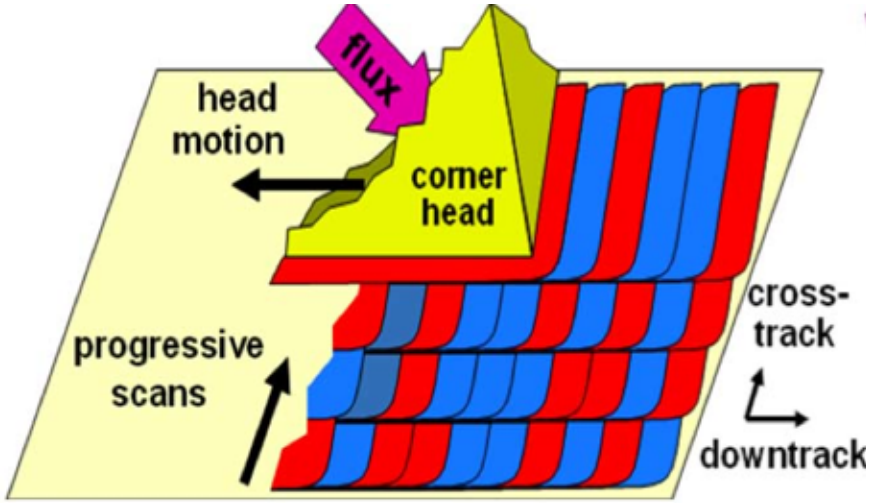
- Heat-Assisted (HAMR)
  - Small bits need high coercivity media to retain orientation
  - High coercivity can't be changed by normal writing
  - Heated media lowers coercivity
  - Include lasers?
- Bit-Patterned (BPM)
  - Small bits retain orientation easier if bits kept apart
  - Pattern media so only write a single dot per bit
  - Tera-dots per sq. inch?





Shingled Magnetic Recording (SMR)

Shingled-Writing



Garth's simple world view

- HAMR, BPMR:
  big changes in fab/assembly
- Shingled-writing does not need big changes

Shingle-writing means

- Partially overwriting tracks, for closer pitch
- Inability to modify one embedded sector without rewriting cross-track neighbors

What About Reading?

Read head is possibly thinner than write head

- If target is 2-3 X density, maybe not too hard

Targeting higher density sees lots of crosstalk

- Signal processing in two dimensions (TDMR)

One approach to TDMR involves gathering signal from 1-2 adjacent tracks on both sides

- Means 3 to 5 revs to read a single sector
- Not likely to be accepted by marketplace

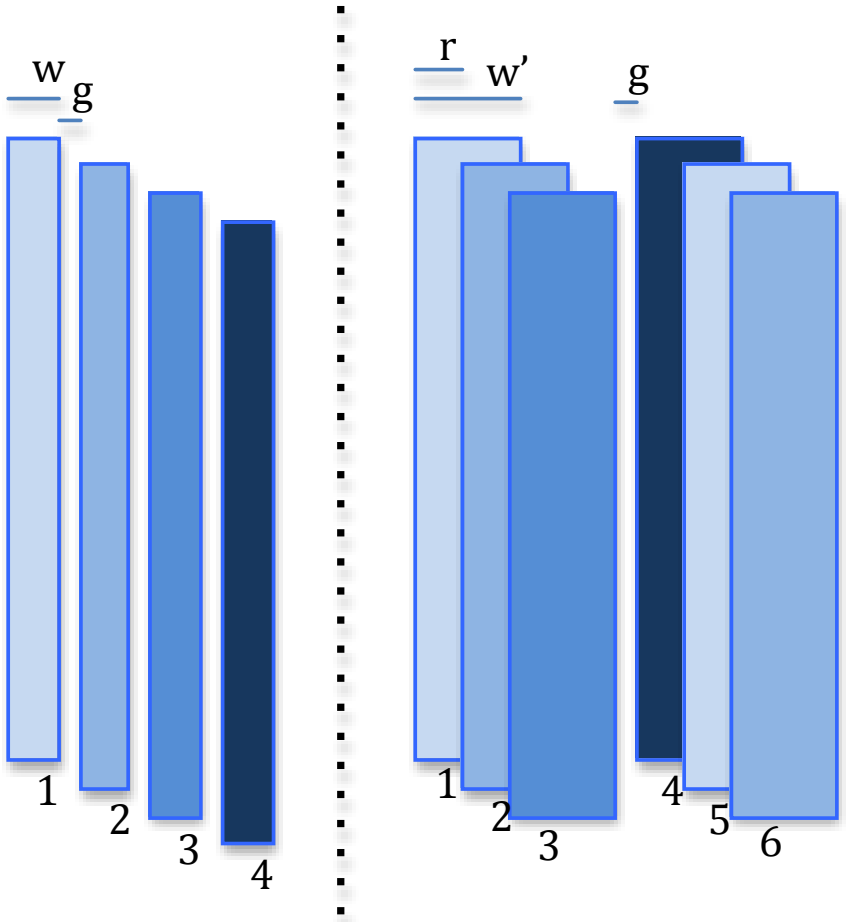
Safe plan is to "see" residual track w/ only 1 head

Geometry Model: Getting a handle on the parameters

Shingled writing: organizational issues

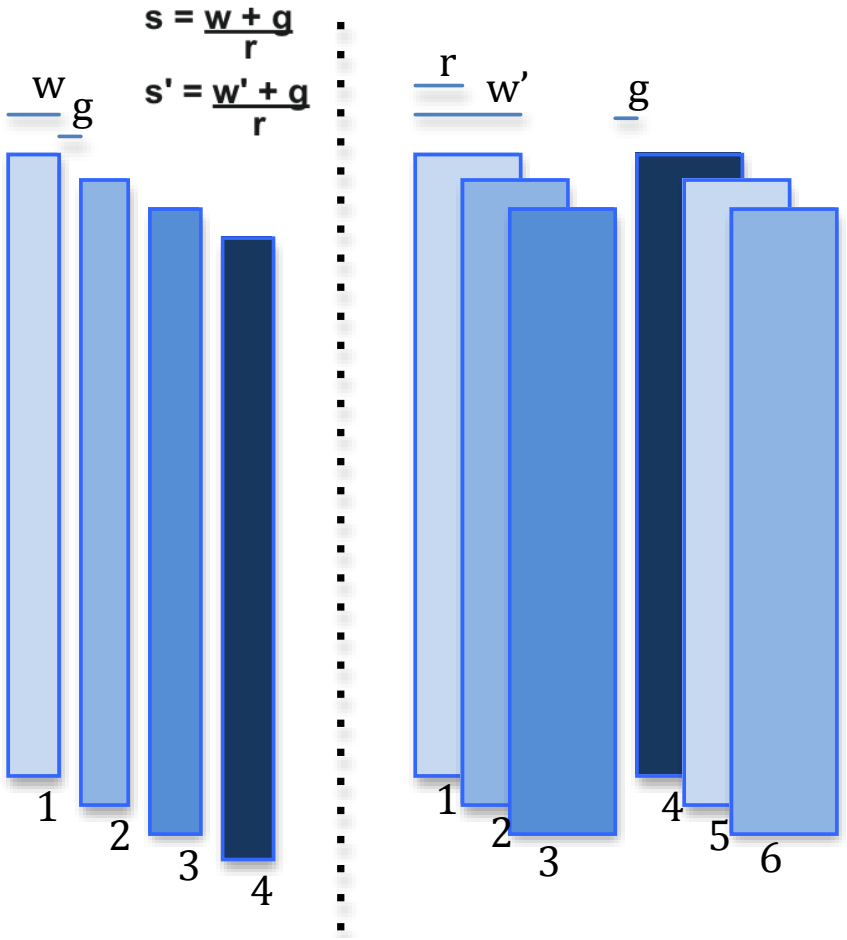
- Reason for doing it: density
  - Shingling projected at 1.5-2.5 X the track density
- Can mix shingled and non-shingled
  - so, e.g., separate sequential from random
  - just lose some of the density gains
- Can break up sets of shingled tracks ("bands")
  - allowing overwrite of individual bands
  - but, they need to be big… like 32 to 256 MB

Simple Geometry Model



- SMR allows wider write heads, w'>w
- SMR reduces gaps, g, per track to per band (B tracks)
- Residual (readable) track width (r) after overlapping is a key factor
- A fraction of tracks not shingled, f, allows some random sector writing

Simple Geometry Model

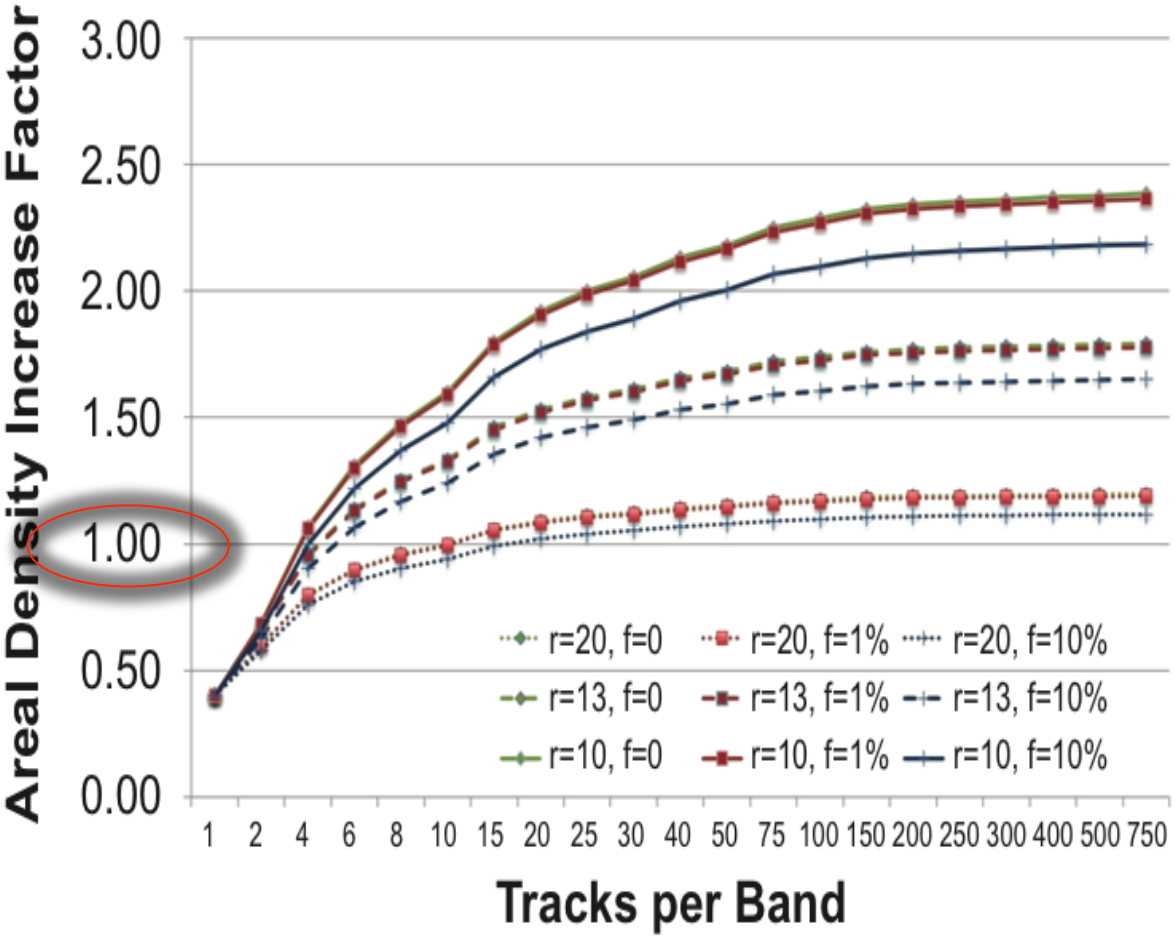


$$\text{Areal Density Increase Factor} = S\left(\frac{f}{s'} + \frac{(1-f)B}{s' + B - 1}\right)$$

- SMR allows wider write heads, w'>w
- SMR reduces gaps, g, per track to per band (B tracks)
- Residual (readable) track width (r) after overlapping is a key factor
- A fraction of tracks not shingled, f, allows some random sector writing
- SMR increase in areal density given by simple model

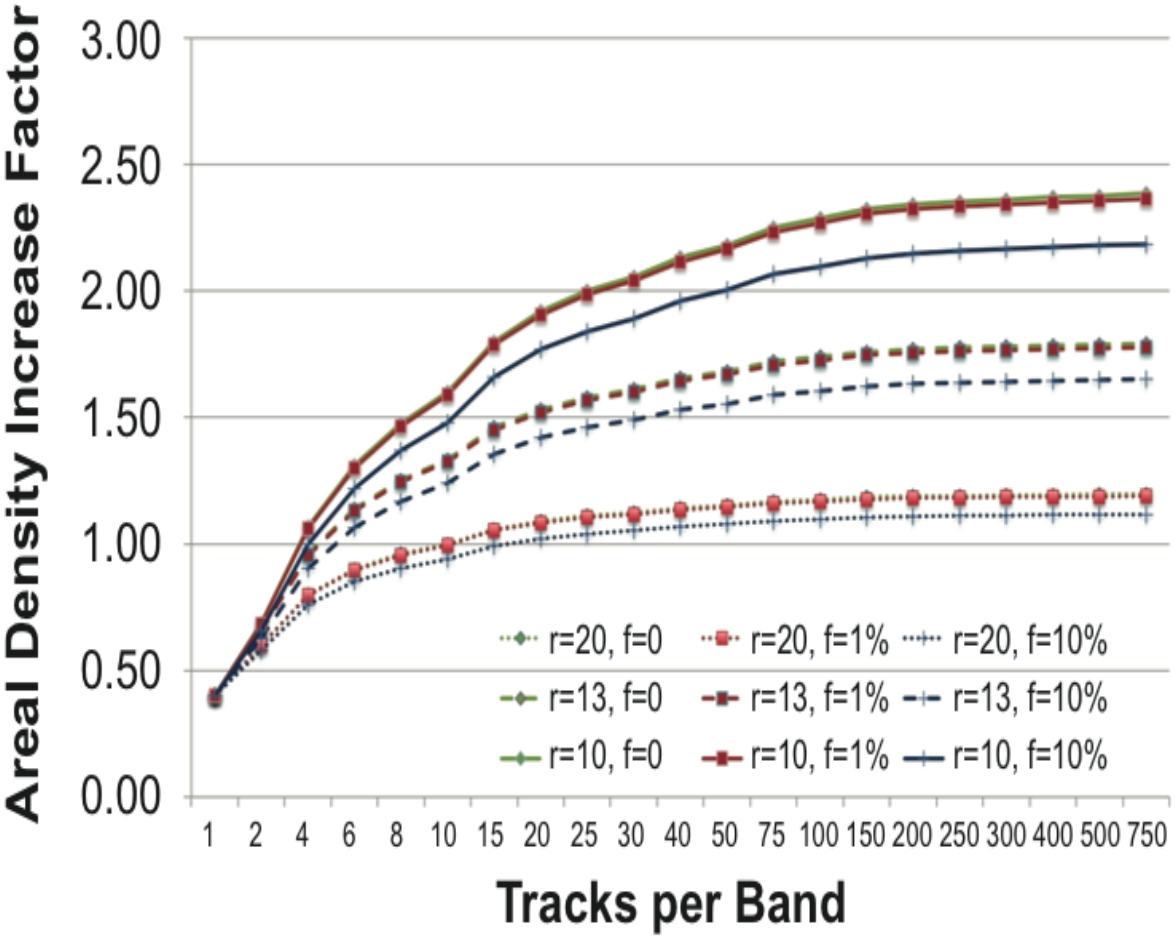
Areal Density Favors Large Bands

Eg. w=25, g=5, w'=70, r=10,13,20 nm, f=0%,1%,10%



Areal Density Favors Large Bands

Eg. w=25, g=5, w'=70, r=10,13,20 nm, f=0%,1%,10%



- 1% unshingled is affordable
  - 10% if r<w
- small B bad news
- r~=w needs large B (~100+)
- r<w allows smallish B (~10)
  - But not soon ….

Systems should plan for large bands

Coping with SMR at the system level

Same Problem for Flash

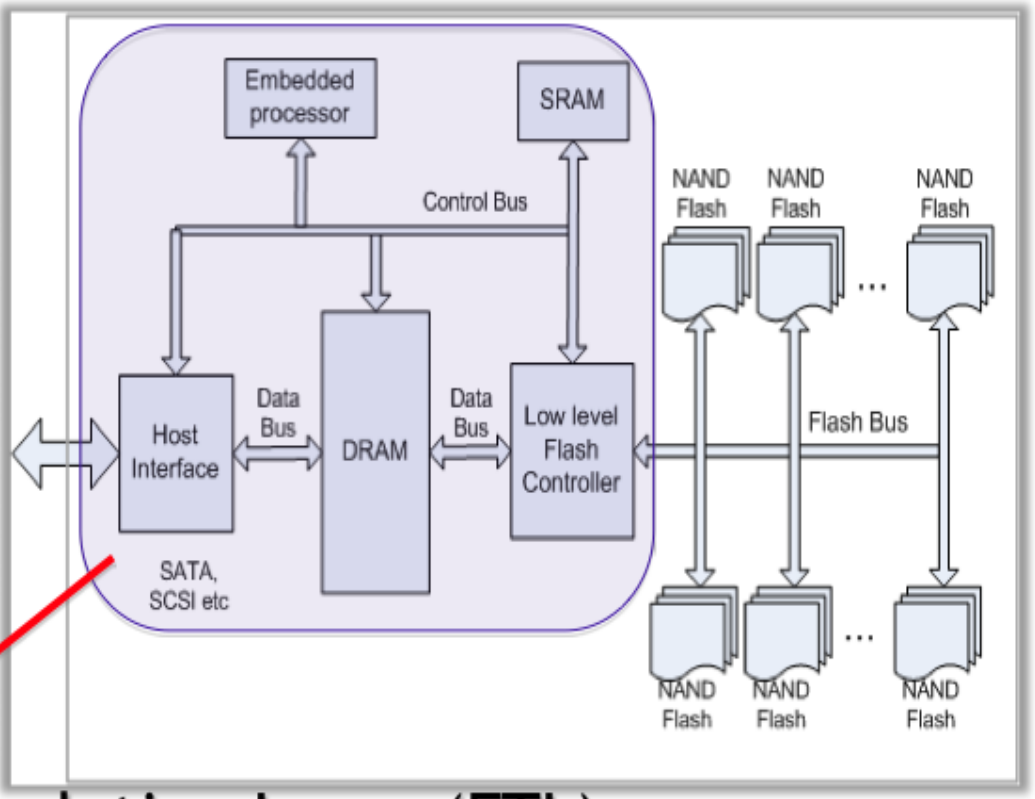
Flash SSD organized as “bands” of “sectors”

Must pre-erase band before programming data

Hide erase in FTL

Simple products
rewrite band
on all writes

Smart products
remap LBN
dynamically



Flash Translation Layer (FTL)

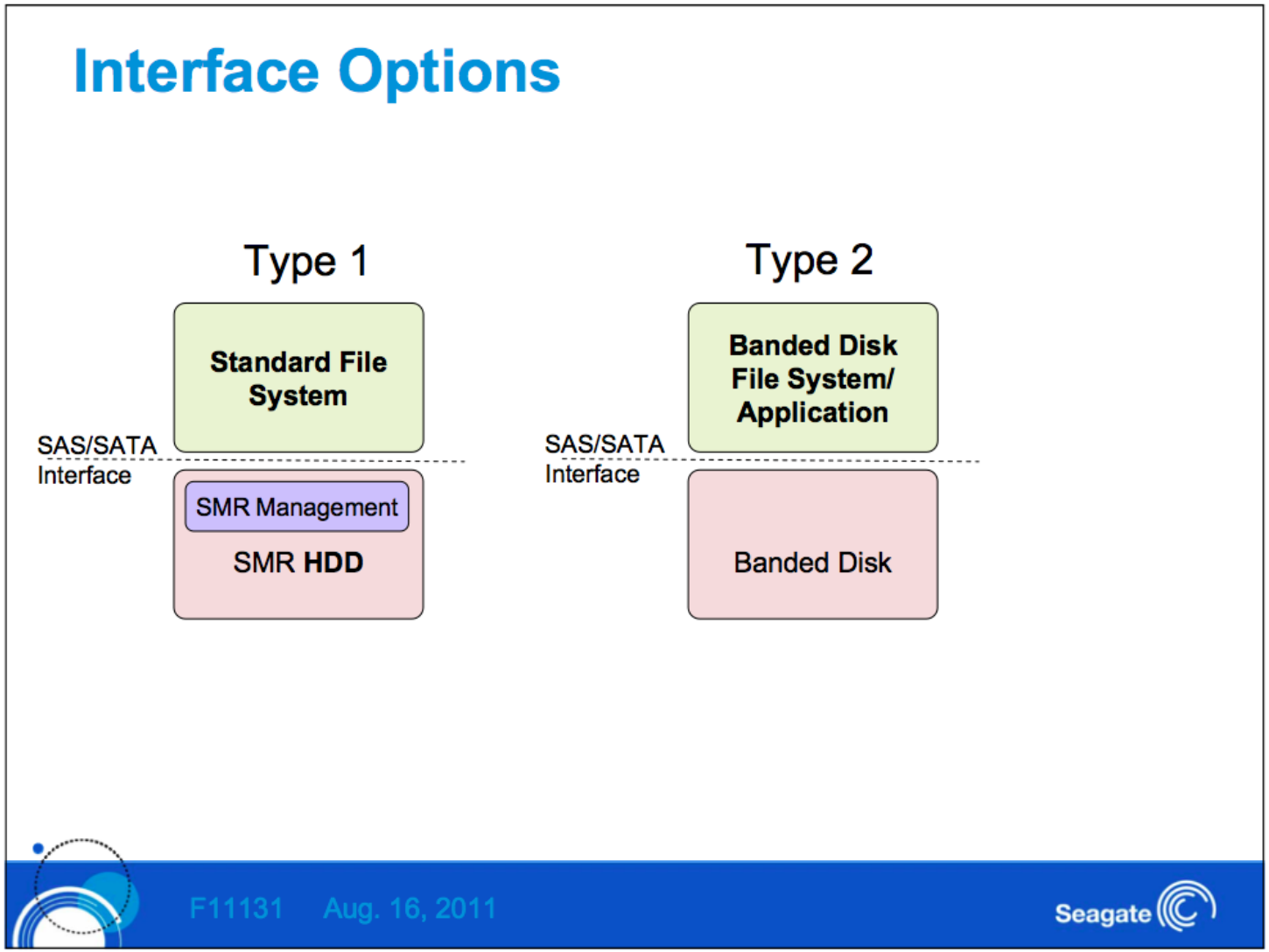
Transparent STL/FTL approach

- Shingled disks implement “translation”
  - Same types of algorithms as Flash
  - Can hire ex-staff of flash industry to jumpstart
  - Data will be correct using existing codes
- Not performance transparent
  - Erase block: 100-1000 X bigger
  - Read-erase-write: 1000-10000 X longer
  - Sure to exceed long tolerable latency thresholds
- Not cost transparent
  - Disk margins < flash margins
  - Yet disk STL needs more resources

Explicitly non-transparent SMR interface

- Define an interface exposing key differences
  - Bands, non-shingled regions, trim, …
- Modify systems software to avoid, minimize read-modify-write
  - Log-structured files systems 20 years old
  - STL-like technology not costly in host
  - Cloud storage writes in 64 MB chunks (HDFS)
  - Flash, PCM, etc may be available to host

A Standards Process is Starting in T13



Shingled Disk Write is really Append

Banded Drives

- Banded Devices
  - Drive divided up into a single Random-write band & multiple sequential write bands
  - All bands are Random-read
  - RD/WR commands address data using a Band # plus LBA offset (RBA) into the band
    - In sequential-write bands the drive always writes to the next sequential block
    - Drives manage band write pointers across power cycles and resets
  - Bands can be 'linked' using Manage Bands command
    - Linked Bands allow RD/WR commands to span multiple bands
    - Links can be changed dynamically by system to manage the user data 'space'
  - Bands are not necessarily aligned to head and media boundaries
  - SCSI Reserve/Release commands supported independently on each band
  - Encryption keys can be aligned with bands enabling independent cryptographic erasure of bands

F11131 Aug. 16, 2011 Seagate

Proposal Applies to Non-Shingled too

Banded Drives

- Concept of a Banded disk is driven by two requirements
  - Drives are getting larger and becoming harder to manage
  - SMR and SSD have unique write requirements that might fit well with a banded disk
- Banded Command Set Approach
  - Supports SMR, SSD, Hybrid and Conventional recording drives
  - Conventional drives would have multiple random-write bands
  - Banded drives would have at least one single random-write band and may have multiple sequential-write access bands
  - Both SCSI and ATA command sets supported
  - Command set changes are focused on RD/WR commands and a few new supporting commands, many existing commands remain unchanged

F11131 Aug. 16, 2011 Seagate

Closing

- Disks are evolving
  - Disk bigots deny tape & flash bigots deny disk
  - But cost & capacity demands prohibit euthanasia
  - Storage hierarchy just gets deeper
- One leading disk evolution overlaps tracks
  - Shingled magnetic recording
  - New interfaces & changes in disk software
  - Trad'l performance projections IFF append only
  - Migration problem is same as disks today